NIH Public Access
Author Manuscript
J R Stat Soc Series B Stat Methodol. Author manuscript; available in PMC 2014 January 01.

Published in final edited form as:
J R Stat Soc Series B Stat Methodol. 2013 January 1; 75(1): 185–206. doi:10.1111/j.1467-9868.2012.01052.x.

Robust estimation for homoscedastic regression in the secondary analysis of case–control data

Jiawei Wei,
Texas A&M University, College Station, USA

Raymond J. Carroll,
Texas A&M University, College Station, USA

Ursula U. Müller,
Texas A&M University, College Station, USA

Ingrid Van Keilegom, and
Université catholique de Louvain, Louvain-la-Neuve, Belgium, and Tilburg University, The Netherlands

Nilanjan Chatterjee
National Cancer Institute, Rockville, USA

Summary

Primary analysis of case–control studies focuses on the relationship between disease D and a set of covariates of interest (Y, X). A secondary application of the case–control study, which is often invoked in modern genetic epidemiologic association studies, is to investigate the interrelationship between the covariates themselves. The task is complicated owing to the case–control sampling, where the regression of Y on X is different from what it is in the population. Previous work has assumed a parametric distribution for Y given X and derived semiparametric efficient estimation and inference without any distributional assumptions about X. We take up the issue of estimation of a regression function when Y given X follows a homoscedastic regression model, but otherwise the distribution of Y is unspecified. The semiparametric efficient approaches can be used to construct semiparametric efficient estimates, but they suffer from a lack of robustness to the assumed model for Y given X. We take an entirely different approach. We show how to estimate the regression parameters consistently even if the assumed model for Y given X is incorrect, and thus the estimates are model robust. For this we make the assumption that the disease rate is known or well estimated. The assumption can be dropped when the disease is rare, which is typically so for most case–control studies, and the estimation algorithm simplifies. Simulations and empirical examples are used to illustrate the approach.

Keywords

Biased samples; Homoscedastic regression; Secondary data; Secondary phenotypes; Semiparametric inference; Two-stage samples

---

© 2012 Royal Statistical Society

Address for correspondence: Raymond J. Carroll, Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, USA. carroll@stat.tamu.edu.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:
'Supplemental material for Robust estimation for homoscedastic regression in the secondary analysis of case-control data'.

1. Introduction

Case–control designs are popularly used for studying risk factors for rare diseases, such as cancers. Under this design, a fixed number of ‘cases’ and ‘controls’, i.e. subjects with and without the disease of interest, are sampled from an underlying base population. Data on various covariates on the subjects are then collected in a retrospective fashion so that they reflect history before the disease. The standard method for primary analysis of case–control data involves logistic regression modelling of the disease outcome as a function of the covariates of interest. It is well known that prospective logistic regression analysis for case–control data is efficient under a semiparametric framework that allows the ‘nuisance’ distribution of the underlying covariates to be unspecified (Prentice and Pyke, 1979).

Epidemiologic researchers popularly use controls from case–control studies to examine the interrelationship between certain covariates themselves. Such secondary analysis of case–control studies has received increasing attention in genetic epidemiologic studies, where it is often of interest to investigate the effect of genetic susceptibility, such as single-nucleotide polymorphism (SNP) genotypes, not only on the primary disease outcome, but also on various secondary factors, such as smoking habits, that may themselves be associated with the disease of interest. For such secondary analysis, use of only controls is generally considered a model robust approach since, when the disease is rare, the relationship between covariates in the controls should reflect that of the underlying population without any further model assumptions. It is, however, recognized that inclusion of cases in such analysis can increase efficiency, provided that appropriate adjustment can be made to account for non-random ascertainment in case–control sampling. Li *et al.* (2010), for example, reported that, if two binary covariates have no interaction with the risk of the disease on a logistic scale, then the association between the factors in the cases remains the same as that for the underlying population. Therefore in such a setting inclusion of cases can increase the efficiency of the secondary analysis.

In this paper, our goal is to develop an approach to secondary association analysis for a continuous covariate, say Y, in a case–control study setting so that both cases and controls can be used to increase efficiency and yet the resulting inference is model robust to distributional assumptions about the covariates. Suppose that data are originally collected from a case–control study of a relatively rare disease. Let D be disease status, with $D = 1$ denoting a case and $D = 0$ denoting a control. Suppose also that D is to be modelled by a vector of random covariates (Y, X), where Y is univariate and X is potentially multivariate, by using a standard logistic regression formulation. Consider here the homoscedastic regression model

$$Y=\alpha_{\text{true}}+\mu(X,\beta_{\text{true}})+\varepsilon, \quad (1)$$

where α_{true} is an intercept and $\mu(\cdot)$ is a known function, and where ε has mean 0 and is independent of X, but its distribution is otherwise not specified.

To estimate $(\alpha_{\text{true}}, \beta_{\text{true}})$, we cannot simply ignore the case–control sampling scheme and use the data *as they are*, because, if Y is a predictor of disease status D, the sampling is biased and in the case–control sample model (1) will not hold.

This paper is organized as follows. In Section 2, we describe recent work on case–control studies that allows efficient estimation if the distribution of Y given X is specified up to parameters. Although the solution is elegant, it suffers from the fact that the resulting estimate may be biased if the hypothesized distribution for Y given X is misspecified.

Section 3 takes an entirely different approach to the basic general problem and describes a simple method that is robust to misspecification of the distribution of Y given X. In Section 4 we describe extensions to cases that the disease rate in the population is known or well estimated from a disease registry or as part of an on-going cohort, and to the case of stratified or frequency-matched studies. Section 5 presents a series of simulation studies, whereas Section 6 presents analysis of an epidemiological data set. Concluding remarks are in Section 7. Technical details are given in Appendix A and Appendix B.

2. Efficient parametric estimation and robustness

2.1. Framework

In this section we outline recent work on efficient estimation for case–control studies when the distribution of Y given X is specified up to a finite dimensional parameter vector. We start with a logistic regression model underlying the case–control analysis, so that $\mathrm{pr}(D = 1|Y, X) = H\{\theta_0 + m(Y, X, \theta_1)\}$, where $H(\cdot)$ is the logistic distribution function and $m(\cdot)$ is an arbitrary known function with unknown parameter vector θ_1. For $d = 0, 1$, let $\pi_d = \mathrm{pr}(D = d)$, the probability that $D = d$ in the population, and suppose that there are n_1 cases with $D = 1$ and n_0 controls with $D = 0$. We write $n = n_0 + n_1$ and introduce the parameter $\kappa = \theta_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$. This reparameterization has the advantage that we can identify κ and θ_1 from a logistic regression analysis of D on (Y, X), although we cannot identify θ_0 (Prentice and Pyke, 1979; Chatterjee and Carroll, 2005) from such logistic regression alone.

In the parametric framework the conditional distribution of Y given X is modelled as $f_\varepsilon\{y - \alpha - \mu(x, \beta), \zeta\}$, where ζ is a finite dimensional nuisance parameter. If in the population Y given X is normally distributed, then $\zeta = \mathrm{var}(\varepsilon)$.

2.2. Population-based case–control studies and notation

Our explicit theoretical and asymptotic results are based on population-based case–control studies, i.e. studies in which random samples of (Y, X) are taken separately for $D = 1$ and $D = 0$. We shall refer to these simply as case–control studies. Some case–control studies use a form of stratification, which is sometimes called frequency matching, e.g. a population-based case–control study for each of a number of age ranges and the same number of cases and controls in each age group. With some notation and the inclusion of these strata in the logistic risk model and in the model for Y given X, our results are easily extended to such sampling; see Section 4.

We assume a logistic model for $\mathrm{pr}(D = 1|Y, X)$ as

$$\mathrm{pr}(D=1|Y,X)=H\{\theta_0+m(Y,X,\theta_1)\}=\frac{\exp\{\theta_0+m(Y,X,\theta_1)\}}{1+\exp\{\theta_0+m(Y,X,\theta_1)\}}. \quad (2)$$

Our technical assumptions are assumptions 1–4 in Appendix B.1.

We also mention two important calculations. The density f_X of X in the population can be written as

$$f_X(x)=\pi_1 f_{\mathrm{case}}(x)+\pi_0 f_{\mathrm{cont}}(x), \quad (3)$$

with (π_0, π_1) defined in Section 2.1, and where $f_{\mathrm{cont}}(x)$ and $f_{\mathrm{case}}(x)$ represent the density of X given $D = 0$ and $D = 1$ respectively. Since this is a case–control sampling scheme, all expectations are conditional on $D_1, \ldots, D_n$. Define $R(\beta) = Y - \mu(X, \beta)$ and $R_i(\beta) = Y_i - \mu(X_i, \beta)$. For an arbitrary function G,

$$\begin{aligned} &\mathrm{E}\left[n^{-1}\sum_{i=1}^{n}G\{R_i(\beta),X_i,D_i\}\right] \\ =&E\left(E\left[n^{-1}\sum_{i=1}^{n}G\{R_i(\beta),X_i,D_i\}|D_1,\ldots,D_n\right]\right)=n^{-1}\sum_{i=1}^{n}E(E[G\{R_i(\beta),X_i,D_i\}|D_i]) \qquad (4)\\ =&\sum_{d=0}^{1}(n_d/n)E[G\{R(\beta),X,d\}|D=d], \end{aligned}$$

the second and last steps following because (Y, X) are independent and identically distributed given D in the case–control sampling scheme.

2.3. Prior results and robustness

For the case–control studies that were described above, Jiang *et al.* (2006), Chen *et al.* (2008) and Lin and Zeng (2009) derived the efficient profile likelihood (in the sense that its score for β is an efficient score function), Lin and Zeng (2009) noting importantly that it can be used in our context. See also Monsees *et al.* (2009). Write $\Omega = (\kappa, \theta_1, \theta_0)$. The joint density of (D, Y, X) is

$$f_X(x)f_\varepsilon\{y-\alpha-\mu(x,\beta),\zeta\}\frac{\exp[d\{\theta_0+m(y,x,\theta_1)\}]}{1+\exp\{\theta_0+m(y,x,\theta_1)\}}.$$

Let

$$g(d,y,x,\Omega,\alpha,\beta,\zeta)=f_\varepsilon\{y-\alpha-\mu(x,\beta),\zeta\}\exp[d\{\kappa+m(y,x,\theta_1)\}][1+\exp\{\theta_0+m(y,x,\theta_1)\}]^{-1}.$$

The semiparametric efficient retrospective profile likelihood for β that makes no assumptions about the distribution of X when the distribution of Y given X is specified is

$$\mathscr{L}_{\mathrm{par}}(D,Y,X,\Omega,\alpha,\beta,\zeta)=\frac{g(D,Y,X,\Omega,\alpha,\beta,\zeta)}{\sum_{d=0}^{1}\int g(d,t,X,\Omega,\alpha,\beta,\zeta)\mathrm{d}t}.$$

Taking logarithms, summing over the observed data and then maximizing in the parameters yields semiparametric efficient inference.

A difficulty arises, however, if the density $f_\varepsilon(\cdot)$ of ε is not specified properly. To see what happens, consider the score for β. Define $L_{\mathrm{par}}(y, x, \alpha, \beta, \zeta) = \partial\log[f_\varepsilon\{y-\alpha-\mu(x,\beta),\zeta\}]/\partial\beta$. Then the score for β is

$$\mathscr{K}_{\mathrm{par}}(D,Y,X,\Omega,\alpha,\beta,\zeta)=\frac{\partial\log\{\mathscr{L}_{\mathrm{par}}(D,Y,X,\Omega,\alpha,\beta,\zeta)\}}{\partial\beta}=L_{\mathrm{par}}(Y,X,\alpha,\beta,\zeta)-\frac{\int\sum_{d=0}^{1}L_{\mathrm{par}}(t,X,\alpha,\beta,\zeta)g(d,t,X,\Omega,\alpha,\beta,\zeta)\mathrm{d}t}{\int\sum_{d=0}^{1}g(d,t,X,\Omega,\alpha,\beta,\zeta)\mathrm{d}t}. \qquad (5)$$

Because $\mathscr{L}_{\mathrm{par}}(\cdot)$ is a legitimate semiparametric profile likelihood, when summed over the case–control data and evaluated at the true parameters, score (5) has mean 0. However, score (5), when evaluated at the true parameter values, only has mean 0 in general if the density $f_\varepsilon(\cdot)$ of ε is specified properly, i.e. the approach is not always model robust; see Section 5 for numerical evidence. This motivates our search for a robust estimation method, which is a topic that we take up in the next section.

3. Model robust estimation

3.1. Preliminaries

In this section we assume the same framework as in the previous section, with the exception that f_ϵ is now unknown. We pursue a sequential approach to derive an estimating equation for the parameters that determine the regression function.

- **a.** Estimate the true logistic regression parameters κ and θ_1 by ordinary logistic regression of D on (Y, X). This can be done legitimately because it is known that ordinary logistic regression in a case–control study consistently estimates κ and θ_1 (Prentice and Pyke, 1979; Chatterjee and Carroll, 2005). Denote the estimators by $\hat{\kappa}$ and $\hat{\theta}_1$.We also suppose that we have a consistent estimator of θ_0. This estimator can, for example, be the solution of the equation

  $$\pi_1=\pi_1 n_1^{-1}\sum_{i=1}^{n} D_i H\{\theta_0+m(Y_i,X_i,\hat{\theta}_1)\}+\pi_0 n_0^{-1}\sum_{i=1}^{n}(1-D_i)H\{\theta_0+m(Y_i,X_i,\hat{\theta}_1)\}, \quad (6)$$

  when the disease rate π_1 in the population is known or well estimated, either from a disease registry or from an underlying cohort from which the cases and controls are sampled. Equation (6) leads to a consistent estimator of θ_0, since for any function $g(y, x)$ we can estimate $\int g(y, x)\, f_{YX}(y, x)\, \mathrm{d}y\mathrm{d}x$ unbiasedly by

  $$\sum_{d=0}^{1}\sum_{i=1}^{n}(\pi_d/n_d)I(D_i=d)g(Y_i,X_i).$$

  Call the resulting estimator $\hat{\theta}_0$ and denote $\hat{\Omega} = (\hat{\kappa}, \hat{\theta}_1, \hat{\theta}_0)$.

- **b.** Use a score function for β that would be an appropriate score function if the (Y, X) data arose from random sampling. Define $R(\beta) = Y - \mu(X, \beta)$. Then the simplest such score function is that from ordinary least squares, which is obtained by differentiating $\{Y - \alpha - \mu(X, \beta)\}^2$ with respect to β. This yields the score function

  $$L\{R(\beta),X,\alpha,\beta\}=\mu_\beta(X,\beta)\{R(\beta)-\alpha\}, \quad (7)$$

  where the subscript means differentiation with respect to β.

- **c.** Score (7) will not have mean 0 in the case–control sampling scheme, so we adjust it so that it has mean 0 in general.

- **d.** For technical reasons that are described later, estimation of α_{true} must be done via an auxiliary equation depending on the current values, which we generically call $\hat{\alpha}(\beta, \Omega)$, which replaces α in score (7); see below for the definition.

- **e.** Solve the adjusted score equation to estimate β_{true} and hence α_{true}. Good starting values for β can be obtained by least squares regression among the controls.

Remark 1. The score function (7) is not the only one possible; for example, we could instead allow for robustness against outliers by replacing function (7) by the estimating equation of an *M*-estimator (Huber, 1981; Anderson, 2008).

3.2. Estimation algorithm

The development of our methodology is somewhat involved. Here we simply state our proposal, with its development given in Sections 3.3–3.5. As before, define $R(\beta) = Y - \mu(X,$

β).Remember that estimation of α_{true} must be done by using an auxiliary equation; see equation (8) directly below. Define

$$\mathscr{K}(R_i(\beta), x, \beta, \Omega)=\frac{1+\exp[\kappa+m\{R_i(\beta)+\mu(x,\beta), x, \theta_1\}]}{1+\exp[\theta_0+m\{R_i(\beta)+\mu(x,\beta), x, \theta_1\}]}.$$

For given (β, Ω), the estimator of α_{true} is justified in Section 3.5 and given by

$$\hat{\alpha}(\beta,\Omega)=\frac{n^{-1}\sum_{i=1}^{n}R_i(\beta)\left[\sum_{d=0}^{1}(\pi_d/n_d)\sum_{j=1}^{n}I(D_j=d)\mathscr{K}\{R_i(\beta),X_j,\beta,\Omega\}\right]^{-1}}{n^{-1}\sum_{i=1}^{n}\left[\sum_{d=0}^{1}(\pi_d/n_d)\sum_{j=1}^{n}I(D_j=d)\mathscr{K}\{R_i(\beta),X_j,\beta,\Omega\}\right]^{-1}}=\frac{n^{-1}\sum_{i=1}^{n}R_i(\beta)\left[n^{-1}\sum_{j=1}^{n}\tilde{\mathscr{K}}\{R_i(\beta),X_j,\beta,\Omega,D_j\}\right]^{-1}}{n^{-1}\sum_{i=1}^{n}\left[n^{-1}\sum_{j=1}^{n}\tilde{\mathscr{K}}\{R_i(\beta),X_j,\beta,\Omega,D_j\}\right]^{-1}}, \quad (8)$$

where

$$\tilde{\mathscr{K}}\{R_i(\beta),X_j,\beta,\Omega,D_j\}=\sum_{d=0}^{1}(n\pi_d/n_d)I(D_j=d)\mathscr{K}\{R_i(\beta),X_j,\beta,\Omega\}.$$

Let $\mu_\beta(x, \beta) = \partial\mu(x, \beta)/\partial\beta$ and let $L\{R(\beta), X, \alpha, \beta\}$ be as in equation (7). Then define

$$\hat{\mathrm{Q}}_{n,\text{est}}(\beta,\Omega)=n^{-1/2}\sum_{i=1}^{n}\left[L\{R_i(\beta),X_i,\hat{\alpha}(\beta,\Omega),\beta\}-\frac{n^{-1}\sum_{j=1}^{n}L\{R_i(\beta),X_j,\hat{\alpha}(\beta,\Omega),\beta\}\tilde{\mathscr{K}}\{R_i(\beta),X_j,\beta,\Omega,D_j\}}{n^{-1}\sum_{j=1}^{n}\tilde{\mathscr{K}}\{R_i(\beta),X_j,\beta,\Omega,D_j\}}\right]. \quad (9)$$

Our algorithm then is as follows.

- **a.** Estimate $(\kappa, \theta_1)^{\mathrm{T}}$ by $(\hat{\kappa}, \hat{\theta}^1)^{\mathrm{T}}$, the logistic regression estimates of D on (Y, X). As described previously, this is known to produce consistent estimates of $(\kappa_{\text{true}}, \theta_{1,\text{true}})^{\mathrm{T}}$. Estimate θ_0 as explained in Section 3.1. This leads to an estimator $\hat{\Omega}$ of Ω_{true}.
- **b.** Solve $0 = \hat{Q}_{n,\text{est}}(\beta, \hat{\Omega})$ in β to obtain the estimate $\hat{\beta}$.

In the next few subsections, we describe how we obtained equation (9), and at the end we describe the asymptotic distribution theory.

3.3. Development of the score when f_X and α_{true} are known

3.3.1. Adjusting score (7)—We first describe how to proceed when the intercept α_{true}, the density $f_X(\cdot)$ of X in the population, and $f_\epsilon(t-\alpha_{\text{true}})$, the density of $Y-\mu(X, \beta_{\text{true}})$ in the population, are all known; they are not and we shall show how to remove these restrictions in subsequent sections.

The approach is to start with the estimating function (7), which, when summed over the data, does not have mean 0 at the true parameters because of the case–control sampling scheme, i.e. $E[\Sigma_{i=1}^{n}L\{R_i(\beta_{\text{true}}), X_i, \alpha_{\text{true}}, \beta_{\text{true}}\}|D_i] \neq 0$, in general. Thus, we need to correct $n^{-1}\Sigma_{i=1}^{n}L\{R_i(\beta), X_i, \alpha, \beta\}$ so that it does have mean 0 in the case–control sampling scheme, where expectations are computed as in equation (4). In the on-line supplemental material, we show how to follow the approach of Chen *et al.* (2009), section 2.3.3, to develop the adjusted estimating function

$$L\{R(\beta), X, \alpha_{\text{true}}, \beta\} - \frac{\int L(t, x, \alpha_{\text{true}}, \beta)\mathscr{K}(t, x, \beta, \Omega) f_{\varepsilon}(t-\alpha_{\text{true}}) f_X(x)\mathrm{d}t\mathrm{d}x}{\int \mathscr{K}(t, x, \beta, \Omega) f_{\varepsilon}(t-\alpha_{\text{true}}) f_X(x)\mathrm{d}t\mathrm{d}x}. \quad (10)$$

This is not of much help, since none of $f_\varepsilon(\cdot)$, $f_X(\cdot)$ or α_{true} are known. In subsequent sections we show how to replace these terms by data-estimated quantities, and thus arrive at equation (9).

3.3.2. Replacing the unknown error density—The problem with expression (10) is that we do not know the form of $f_\varepsilon(\cdot)$, so score (10) cannot be implemented. Similarly to Chatterjee and Carroll (2005) and Spinka *et al.* (2005), we therefore replace $f_\varepsilon(\cdot)$ by a non-parametric maximum likelihood estimator. The idea is to take the observed $R_i(\beta) = Y_i - \mu(X_i, \beta)$ as the support, and to maximize the log-likelihood with respect to $\gamma_i = \text{pr}\{R(\beta) = R_i(\beta)\}$, $i = 1, \ldots, n$, subject to $\Sigma_{i=1}^{n}\gamma_i=1$. By Chatterjee and Carroll (2005) and Spinka *et al.* (2005), the resulting estimator for $\text{pr}\{R(\beta) = R_i(\beta)\}$ is

$$p_{\text{est}}\{R_i(\beta), \Omega\} = \frac{\pi_0}{n_0}\left[\int f_X(x)\mathscr{K}\{R_i(\beta), x, \beta, \Omega\}\mathrm{d}x\right]^{-1}. \quad (11)$$

The derivation of equation (11) is given in Appendix A.1. When we make this substitution in expression (10) and sum over the data, the score becomes

$$\sum_{i=1}^{n} L\{R_i(\beta), X_i, \alpha_{\text{true}}, \beta\} - \frac{\sum_{i=1}^{n}\int L\{R_i(\beta), x, \alpha_{\text{true}}, \beta\}\mathscr{K}\{R_i(\beta), x, \beta, \Omega\} p_{\text{est}}\{R_i(\beta), \Omega\} f_X(x)\mathrm{d}x}{n^{-1}\sum_{i=1}^{n}\int \mathscr{K}\{R_i(\beta), x, \beta, \Omega\} p_{\text{est}}\{R_i(\beta), \Omega\} f_X(x)\mathrm{d}x}.$$

Because the denominator of this expression is π_0/n_0, by simple algebra it is readily seen that the normalized score function for estimating β can be defined as

$$0 = Q_n(\alpha_{\text{true}}, \beta, \Omega) = n^{-1/2}\sum_{i=1}^{n}\left[L\{R_i(\beta), X_i, \alpha_{\text{true}}, \beta\} - \frac{\int L\{R_i(\beta), x, \alpha_{\text{true}}, \beta\}\mathscr{K}\{R_i(\beta), x, \beta, \Omega\} f_X(x)\mathrm{d}x}{\int \mathscr{K}\{R_i(\beta), x, \beta, \Omega\} f_X(x)\mathrm{d}x}\right]. \quad (12)$$

In Appendix A.2 we show that the expectation of $Q_n(\alpha_{\text{true}}, \beta, \Omega)$ in the case–control sampling scheme is equal to 0 when evaluated at $(\alpha_{\text{true}}, \beta_{\text{true}}, \Omega_{\text{true}})$, but not for arbitrary (β, Ω). This implies that equation (12) is indeed an unbiased estimating equation in the *case–control sampling* scheme.

3.4. Implementation when f_X is unknown but α_{true} is known

The density or mass function $f_X(\cdot)$ is not known. We estimate the integrals in expression (12) unbiasedly by their sample average over all the observations, so our estimating equation is

$$0 = \hat{Q}_n(\alpha_{\text{true}}, \beta, \Omega) = n^{-1/2}\sum_{i=1}^{n}\left[L\{R_i(\beta), X_i, \alpha_{\text{true}}, \beta\} - \frac{n^{-1}\sum_{j=1}^{n} L\{R_i(\beta), X_j, \alpha_{\text{true}}, \beta\}\tilde{\mathscr{K}}\{R_i(\beta), X_j, \beta, \Omega, D_j\}}{n^{-1}\sum_{j=1}^{n}\tilde{\mathscr{K}}\{R_i(\beta), X_j, \beta, \Omega, D_j\}}\right]. \quad (13)$$

3.5. Implementation when the intercept α_{true} is unknown

One might reasonably think that estimating the intercept is easy; for example, simply supplement the score with the ordinary least squares score for the intercept, so that $L\{R(\beta), X, \alpha, \beta\}=(1, \mu_{\beta}^{\mathrm{T}}(X, \beta))^{\mathrm{T}}\{R(\beta)-\alpha\}$. The problem with this is that the first component of the estimating equation (13) would then be identically 0 and thus will not produce an estimate of the intercept. The reason for this is that the solution (11) was calculated non-parametrically under the assumption that $R(\beta_{\mathrm{true}})$ and X are independent in the population. Since $Y-\alpha_{\mathrm{true}}-\mu(X, \beta_{\mathrm{true}})$ and $Y-\mu(X, \beta_{\mathrm{true}})$ are both independent of X in the population, this means that equation (11) cannot lead to an estimate of the intercept. Hence, an alternative approach is required.

To overcome this problem, we estimate the intercept of $R(\beta)$ by using equation (11), i.e., if $f_X(\cdot)$ were known, then α_{true} could be estimated by

$$\tilde{\alpha}(\beta, \Omega)=\frac{n^{-1}\sum_{i=1}^{n} R_i(\beta) p_{\mathrm{est}}\{R_i(\beta), \Omega\}}{n^{-1}\sum_{i=1}^{n} p_{\mathrm{est}}\{R_i(\beta), \Omega\}}: \quad (14)$$

a quantity that is free of the π_0 that shows up in equation (11). If we then replace the integral in the definition of $p_{\mathrm{est}}(\cdot)$ by its average $n^{-1}\Sigma_{j=1}^{n}\tilde{\mathscr{K}}\{R_i(\beta), X_j, \beta, \Omega, D_j\}$, we obtain exactly expression (8). Making this substitution in equation (13), we obtain equation (9). This completes the derivation of our methodology.

3.6. Distribution theory

The asymptotic distribution of our estimator is given in the following result. We refer to Appendix B.1 for the definition of the functions and matrices that are mentioned below, and for the assumptions 1–4 there under which this result is valid. The proof of this theorem is given in Appendix B.2.

Theorem 1. Let $(\beta, \Omega) = \Theta$, and let Θ_{true} denote its true value. Assume that assumptions 1–4 in Appendix B.1 are valid. Then there is an invertible matrix $\mathscr{M}_{\beta}$ and a function $\Lambda(Y, X, D, \Theta_{\mathrm{true}})$ with the properties that $E\{\Lambda(Y, X, D, \Theta_{\mathrm{true}})|D\} = 0$ and

$$n^{1/2}(\hat{\beta}-\beta_{\mathrm{true}})=-n^{-1/2}\mathscr{M}_{\beta}^{-1}\sum_{i=1}^{n}\Lambda(Y_i, X_i, D_i, \Theta_{\mathrm{true}})+o_p(1).$$

Therefore, there is a matrix Σ, defined in Appendix B.1, such that

$$n^{1/2}(\hat{\beta}-\beta_{\mathrm{true}}) \rightarrow N(0, \Sigma). \quad (15)$$

Estimating the covariance matrix Σ in expression (15) can be accomplished by a plug-in method or by the bootstrap appropriate for case–control sampling (Wang *et al.*, 1997; Buonaccorsi, 2010).

3.7. Inference via bootstrap resampling

In principle, estimating the covariance matrix Σ in expression (15) can be accomplished by a plug-in method, although the particular form of the function $Q_1(\cdot)$ that is defined in Appendix B.1 makes computational speed slow. We have thus chosen to use bootstrap ideas to estimate Σ. Below we explain in detail how this can be done, but the basic idea is that we have random samples from two independent populations, i.e. the cases and the controls, and an estimator that is asymptotically normally distributed.

3.7.1. Bootstrap procedure—Let $(Y_1^*, X_1^*), \ldots, (Y_{n_0}^*, X_{n_0}^*)$ be drawn randomly with replacement from $\{(Y_i, X_i) : D_i = 0\}$, and similarly let $(Y_{n_0+1}^*, X_{n_0+1}^*), \ldots, (Y_n^*, X_n^*)$ be drawn randomly with replacement from $\{(Y_i, X_i) : D_i = 1\}$. This is the method of bootstrap sampling that was suggested by Wang *et al.* (1997) and Buonaccorsi (2010), page 225, and, since the data consist of samples from two independent populations, is the same as in Babu and Singh (1983); see also Lele (1991).

Let $D_i^* = I(i > n_0)$ and $R_i^*(\beta) = Y_i^* - \mu(X_i^*, \beta)$, and define $\hat{\Omega}^*$, $\hat{\alpha}^*(\beta, \Omega)$ and $\hat{Q}^*_{n,\mathrm{est}}(\beta, \Omega)$ in the same way as $\hat{\Omega}$, $\hat{\alpha}(\beta, \Omega)$ in equation (8) and $\hat{Q}_{n,\mathrm{est}}(\beta, \Omega)$ in equation (9), but based on (Y_i^*, X_i^*, D_i^*) instead of (Y_i, X_i, D_i), $i = 1, \ldots, n$.

The bootstrapped estimator $\hat{\beta}^*$ of β is then defined as a solution of

$$0=\hat{Q}^*_{n,\mathrm{est}}(\beta, \hat{\Omega}^*) - \hat{Q}_{n,\mathrm{est}}(\hat{\beta}, \hat{\Omega})=\hat{Q}^*_{n,\mathrm{est}}(\beta, \hat{\Omega}^*)$$

with respect to β. See also Hall and Horowitz (1996), page 897, and Chen *et al.* (2003), where bootstrapping is used and justified in similar contexts.

3.7.2. Bootstrap consistency—To show the consistency of the above bootstrap procedure, we need to show that $n^{1/2}(\hat{\beta}^* - \hat{\beta})$ converges to the same normal limit as the original centred estimator $n^{1/2}(\hat{\beta} - \beta_{\mathrm{true}})$. For this we use the same techniques as in the proof of theorem B in Chen *et al.* (2003), combined with the proof of theorem 1 in Appendix A. More precisely, it can be shown that, under certain regularity conditions, we have that

$$n^{1/2}(\hat{\beta}^* - \hat{\beta}) = -\mathscr{M}_\beta^{-1} n^{-1/2} \sum_{i=1}^{n} \{\Lambda(Y_i^*, X_i^*, D_i^*, \Theta_{\mathrm{true}}) - \Lambda(Y_i, X_i, D_i, \Theta_{\mathrm{true}})\} + o_{p^*}(1),$$

where $o_{p^*}(1)$ has the same meaning as $o_p(1)$, except that the probability is computed under the bootstrap distribution conditional on the original data (Y_i, X_i, D_i), $i = 1, \ldots, n$. From this together with the central limit theorem and theorem 1 the result follows.

4. Extensions

4.1. Rare disease approximations

The method that was defined in Section 3 assumes that $\pi_1 = \mathrm{pr}(D = 1)$ is known. This is typically not the case, so many researchers adopt rare disease approximations (see below for references), where the word 'rare' has no precise definition but is certainly 1% or less. There are at least two ways to proceed in our context. The first is to use the literature, to choose a nominal $\pi_1 \leq 1\%$ and to apply the method in Section 3. In results that are not reported here, this works well in the simulation setting of Section 5. In the literature, most researchers use a different approximation, which is described next and implemented in Section 5. We have not investigated in any detail which approach is preferable.

Let '$\doteq$' denote 'approximately equal'. The estimation procedure simplifies if the disease can be assumed to be *rare*, i.e. if

$$\mathrm{pr}(D=1|Y,X)=\frac{\exp\{\theta_0+m(Y,X,\theta_1)\}}{1+\exp\{\theta_0+m(Y,X,\theta_1)\}}\doteq\exp\{\theta_0+m(Y,X,\theta_1)\},$$

or, equivalently, if $\mathrm{pr}(D=0|\,Y,X)=[1+\exp\{\theta_0+m(Y,X,\theta_1)\}]^{-1}\doteq 1$. This approximation allows us to replace $\mathscr{K}$ in the estimating function (12) by

$$\mathscr{K}^*\{R_i(\beta),x,\beta,\Omega^*\}=1+\exp\left[\kappa+m\{R_i(\beta)+\mu(x,\beta),x,\theta_1\}\right]. \quad (16)$$

In addition, $\Omega=(\kappa,\theta_1,\theta_0)$ in $\mathscr{K}$ is replaced by $\Omega^*=(\kappa,\theta_1)$, which does not depend on θ_0 any-more, and assumption 4 is no longer required since θ_0 is no longer estimated. The proof in Appendix A.2, where we show that the estimating function (12) is unbiased, adapts to the rare disease case in a straightforward way, now using the approximation

$$f_{YX|D=d}(y,x)=\frac{\exp[d\{\theta_0+m(y,x,\theta_1)\}]f_{YX}(y,x)}{[1+\exp\{\theta_0+m(y,x,\theta_1)\}]\pi_d}\doteq\frac{\exp[d\{\theta_0+m(y,x,\theta_1)\}]f_{YX}(y,x)}{\pi_d}.$$

Hence the modified estimating function based on $\mathscr{K}$ is approximately unbiased in the rare disease case.

As in the general case, the rare disease version of the estimating function (12) depends on unknown quantities which must be estimated. The estimation algorithm for the rare disease model is as follows and is explained below. Set

$$\hat{\alpha}^*(\beta,\Omega^*)=\frac{n^{-1}\sum_{i=1}^{n}R_i(\beta)\left[n_0^{-1}\sum_{j=1}^{n}(1-D_j)\mathscr{K}^*\{R_i(\beta),X_j,\beta,\Omega^*\}\right]^{-1}}{n^{-1}\sum_{i=1}^{n}\left[n_0^{-1}\sum_{j=1}^{n}(1-D_j)\mathscr{K}^*\{R_i(\beta),X_j,\beta,\Omega^*\}\right]^{-1}},$$

$$\hat{\mathrm{Q}}^*_{n,\mathrm{est}}(\beta,\Omega^*)=n^{-1/2}\sum_{i=1}^{n}\left[L\{R_i(\beta),X_i,\hat{\alpha}^*(\beta,\Omega^*),\beta\}-\frac{n_0^{-1}\sum_{j=1}^{n}(1-D_j)L\{R_i(\beta),X_j,\hat{\alpha}^*(\beta,\Omega^*),\beta\}\mathscr{S}^*\{R_i(\beta),X_j,\beta,\Omega^*\}}{n_0^{-1}\sum_{j=1}^{n}(1-D_j)\mathscr{K}^*\{R_i(\beta),X_j,\beta,\Omega^*\}}\right].$$

As before, estimate $\Omega^*=(\kappa,\theta_1)$ by the logistic regression estimates of D on (Y,X); then solve $\hat{\mathrm{Q}}^*_{n,\mathrm{est}}(\beta,\hat{\Omega}^*)=0$ with respect to β to obtain $\hat{\beta}$.

The formulae for $\hat{\alpha}^*$ and $\hat{\mathrm{Q}}^*_{n,\mathrm{est}}$ do not contain an average $\tilde{\mathscr{K}}^*$,which could be introduced analogously to the general case where both formulae involve $\tilde{\mathscr{K}}$, and which depends on $\pi_1=P(D=1)$. This is explained as follows: both the estimating function (12) and the estimator p_{est}, which is used to estimate α_{true}, depend on the unknown density f_X. As already explained in Section 2 at equation (3), under the rare disease approximation, f_X can be approximated by f_{cont}, i.e. we can use f_X empirically using only the controls. This has the advantage that we do not need prior knowledge about the typically unknown disease rate π_1. This is in contrast with the general model where we need to know π_1 not only to be able to work with $\tilde{\mathscr{K}}$, but also to obtain a consistent estimator of θ_0.

Because case–control studies are almost inevitably conducted for rare outcomes, the rare disease approximation is natural in most applications. It is also widely used, a very non-exhaustive list of which includes Piegorsch *et al.* (1994), Epstein and Satten (2003), Lin and Zeng (2006), Modan *et al.* (2001), Zhao *et al.* (2003), Kwee *et al.* (2007), Lin and Zeng (2009) and Hu *et al.* (2010).

4.2. Case–control studies with frequency matching

In frequency-matched case–control studies, a few strata are formed based on covariates such as age, and then a population-based case–control study is performed within each stratum. A straightforward approach is to include these matching variables as part of X, to form the estimating function (9) for each stratum and to form a new estimating function as the possibly weighted sum of the estimating functions across the strata. The weights might for example be based on estimates of the size of each stratum in the population. The resulting estimates of $(\alpha_{\text{true}}, \beta_{\text{true}})$ will be asymptotically normally distributed.

5. Simulations

We performed simulation studies both at and away from the Gaussian model. Our simulations indicate that our proposed estimator has small bias and nearly nominal coverage probability in the cases that we examined, whereas an implementation of the parametric approach (see Section 2.3) may suffer from bias and lower coverage probability (Tables 1 and 2). We also show that our method often achieves significant gains in efficiency when compared with the estimator that uses only the controls. The approach that uses all the data but ignores the case–control sampling design suffers from bias and low coverage; see below.

We generated X from a uniform distribution on (0, 1). The logistic regression model is $\text{pr}(D = 1|Y, X) = H(\theta_0 + \theta_y Y + \theta_x X)$, with $\theta_0 = -5.5$, $\theta_y = 0.00, 0.25, 0.50$ and $\theta_x = 1$. The model for Y given X is a linear regression model, $Y = \alpha_{\text{true}} + \beta_{\text{true}} X + \varepsilon$, with $\alpha_{\text{true}} = 0$ and $\beta_{\text{true}} = 1$. We considered two distributions for ε: the standard normal distribution, for which the parametric approach attains the semiparametric efficiency bound, and, for comparison, a standardized gamma distribution with scale parameter 0.4. By equation (2), for $\theta_y = 0.00$, 0.25, 0.50 the rates of disease are approximately 0.007, 0.008 and 0.010. In the first scenario the case–control study has $n_1 = 500$ cases and $n_0 = 500$ controls. In the second scenario we chose $n_0 = n_1 = 150$. We generated 1000 simulated data sets in each setting.

We contrasted four methods. The first uses ordinary linear regression based only on the controls. The second method uses the same approach but is expected to be significantly biased since it is based on the entire data set. The third method is the parametric ('semiparametric efficient') method that assumes normal errors, with standard errors obtained by inverting the Hessian of the log-likelihood. The fourth method is our proposed method, with standard errors estimated by using asymptotic formulae. The third and the fourth method were computed by making the rare disease approximation.

The case $\theta_y = 0.00$ is interesting, because here Y is independent of D given X. Hence all methods should achieve nominal coverage probabilities for estimating β_{true}, which is indeed seen in Table 1. Since, with $\theta_y = 0.00$, all methods are asymptotically valid, the only possibility of seeing a bias is when θ_y is sufficiently 'large'. For this reason, we experimented with the cases $\theta_y = 0.25$ and $\theta_y = 0.50$. Consider $\theta_y = 0.25$ first. Here the approach that uses all the data yields a biased estimator of $\beta_{\text{true}} = 1$, with low coverage probabilities. The 'semiparametric efficient' method that assumes normality still maintains its nominal coverage probabilities. As expected, since it is efficient if the errors are normal, it indeed outperforms the other approaches in this case. For example, for any two methods, say A and B, with estimates $\hat{\beta}_{\text{A}}$ and $\hat{\beta}_{\text{B}}$, the mean-squared error efficiency of method A with respect to method B is $E\{(\hat{\beta}_B - \beta_{true})^2\}/E\{(\hat{\beta}_A - \beta_{true})^2\}$, and its estimated version is computed by replacing expectations by averages across the simulations. The semiparametric efficient method has 13% greater mean-squared error efficiency than our method in the normal case. However, in the gamma case, our method has 43% greater mean-squared error efficiency. It also outperforms the approach that uses only the controls, for both normal and gamma errors: in both cases the mean-squared error efficiency is roughly 70% larger.

Finally, in the case $\theta_Y = 0.50$ with normal regression errors, the semiparametric efficient method that assumes normality maintains its nominal coverage probabilities and has 7% greater mean-squared error efficiency than our method and 82% greater efficiency than using only controls. However, when the errors have a gamma distribution, it suffers from bias, increased variance and loss of coverage, with nominal 90% and 95% coverage actually being 81.8% and 88.4% respectively. Our method retains nominal coverage. The controls-only analysis and our method have roughly equal mean-squared error efficiency which is, in particular, much greater than the mean-squared error efficiency of the semiparametric efficient approach for regression models with normal errors.

6. Empirical example

In this section, we illustrate the methodology in a case–control study of prostate cancer, which was originally designed to investigate the risk of prostate-cancer-associated vitamin D biomarkers and genetic variations in vitamin D metabolism pathways (Ahn *et al.*, 2009). The goal of the current analysis, which includes 749 prostate cancer cases and 781 controls, is to examine whether the genetic variations in the vitamin D receptor influence [25(OH)D], which is a serum level biomarker of vitamin D. In the notation of this paper, D is the prostate cancer case–control status and Y is the level of [25(OH)D]. We investigated three SNPs, rs2238136, rs2254210 and rs2239186, each of which represents an ordinal categorical variable coded as 0, 1 or 2 depending on how many copies of the variant allele a subject carries. In our analysis, X consists of three dummy variables for age groups, along with one of the genetic markers.

The results are given in Table 3. We see in Table 3 that none of the coefficients for the SNP are statistically significant. Thus, neither the traditional control-only nor the proposed method detected any association between the vitamin D receptor gene and [25(OH)D] level. These results are consistent with Chen *et al.* (2009) who noted that, given the downstream role of the vitamin D receptor gene in the vitamin D pathway, it is unlikely that vitamin D receptor polymorphisms could actually influence the level of [25(OH)D]. In spite of a lack of association, it is interesting to observe that the 95% confidence intervals by using our method are much shorter than by using those from the control data only. In terms of mean-squared error efficiency, here estimated as the square of the ratio of the lengths of the confidence intervals, the results for the three SNPs suggest gains in efficiency of 68%, 136% and 125% compared with using only the controls.

7. Discussion

If the disease probability $\mathrm{pr}(D = 1)$ is known, there are simpler methods for our particular setting that allow estimation of β_{true}, based on weighting via equation (3). However, in the common case that $\mathrm{pr}(D = 1)$ is not known, the development in Section 3 leads to two natural rare disease approximations that use all the data and not just the data on the controls; see Section 4.1. It would be interesting to investigate which of these two approximate approaches is preferable in general.

Our simulation results are specific to rare diseases, by which we mean certainly that $\mathrm{pr}(D = 1) \leq 1\%$. Biases will arise as the disease probability increases. In addition, since rare disease approximations do not lead to fully consistent estimation, coverage probability in large samples will suffer, since the bias is fixed whereas the variance decreases with sample size. Finally, the methods are likely to suffer in cases that the *X*-distribution has relatively rare values that are not within the centre of the support of *X*.

Acknowledgments

This paper represents part of the first author's doctoral dissertation at Texas A&M University. Wei and Carroll's research was supported by a grant from the National Cancer Institute (R37-CA057030). Carroll was also supported by award KUS-CI-016-04, made by King Abdullah University of Science and Technology. Chatterjee's research was supported by a gene–environment initiative grant from the National Heart, Lung and Blood Institute (RO1-HL091172-01) and by the Intramural Research Program of the National Cancer Institute. Muller was supported by a National Science Foundation grant (DMS-0907014). Van Keilegom gratefully acknowledges financial support from Interuniversity Attraction Pole research network P6/03 of the Belgian Government (Belgian science policy), and from the European Research Council under the European Community's seventh framework programme (FP7/2007-2013), European Research Council grant agreement 203650.

Appendix A

Some derivations

A.1. Derivation of the error density estimator (11)

The key idea of the approach is to introduce discrete probabilities $\gamma_i = \mathrm{pr}\{R(\beta) = R_i(\beta)\}$, $i = 1, \ldots, n$, which yields

$$\mathrm{pr}(D=d)=\sum_{i=1}^{n}\mathrm{pr}\{D=d|R(\beta)=R_i(\beta)\}\gamma_i,$$

and to work with the maximum likelihood estimates, i.e. with those γ_i that maximize the retrospective log-likelihood

$$\sum_{i=1}^{n}\log[\mathrm{pr}\{R(\beta)=R_i(\beta)|D=D_i\}]=\sum_{i=1}^{n}\log\left[\frac{\mathrm{pr}\{R(\beta)=R_i(\beta)\}\mathrm{pr}\{D=D_i|R(\beta)=R_i(\beta)\}}{\mathrm{pr}(D=D_i)}\right]=\sum_{i=1}^{n}\log\left[\sum_{k=1}^{n}\gamma_k\mathbf{1}\{R_i(\beta)=R_k(\beta)\}\right]+\sum_{i=1}^{n}\log\left[\frac{\mathrm{pr}\{D=D_i|R(\beta)=R_i(\beta)\}}{\sum_{k=1}^{n}\mathrm{pr}\{D=D_i|R(\beta)=R_k(\beta)\}\gamma_k}\right].$$

Taking the derivative with respect to γ_k, $k = 1, \ldots, n$, gives

$$\frac{\sum_{i=1}^{n}I\{R_i(\beta)=R_k(\beta)\}}{\gamma_k}-\sum_{i=1}^{n}\frac{\mathrm{pr}\{D=D_i|R(\beta)=R_k(\beta)\}}{\sum_{k=1}^{n}\mathrm{pr}\{D=D_i|R(\beta)=R_k(\beta)\}\gamma_k}=\gamma_k^{-1}-\sum_{i=1}^{n}\frac{\mathrm{pr}\{D=D_i|R(\beta)=R_k(\beta)\}}{\mathrm{pr}(D=D_i)}=\gamma_k^{-1}-\sum_{d=0}^{1}\mathrm{pr}\{D=d|R(\beta)=R_k(\beta)\}\frac{n_d}{\pi_d}.$$

Now set this equal to 0 to obtain

$$\gamma_k=\left[\sum_{d=0}^{1}\mathrm{pr}\{D=d|R(\beta)=R_k(\beta)\}\frac{n_d}{\pi_d}\right]^{-1}=\left[\int\sum_{d=0}^{1}\mathrm{pr}\{D=d|R(\beta)=R_k(\beta),X=x\}f_X(x)dx\frac{n_d}{\pi_d}\right]^{-1}.$$

By definition of $\mathcal{K}$ using that

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

$$\frac{n_0}{\pi_0}+\frac{n_1}{\pi_1}\exp\{\theta_0+m(y,x,\theta_1)\}=\frac{n_0}{\pi_0}[1+\exp\{\kappa+m(y,x,\theta_1)\}],$$

this is the desired formula (11).

A.2. Unbiasedness of estimation function (12)

All calculations of expectations here will be based on the precise definition of expectations in a case–control sampling scheme; see equation (4). Let $(\beta_{\mathrm{true}}, \Omega_{\mathrm{true}})$ be the true parameter, β an arbitrary value and $\tau(x, \beta, \beta_{\mathrm{true}}) = \mu(x, \beta_{\mathrm{true}}) - \mu(x, \beta)$. To derive the conditional density given the disease state we use the fact that we assume a logistic model, $\mathrm{pr}(D = 1| Y, X) = H\{\theta_0 + m(Y, X, \theta_1)\}$, with $H(x)$ the logistic distribution function, for which

$$H\{\theta_0+m(Y,X,\theta_1)\}=[1-H\{\theta_0+m(Y,X,\theta_1)\}]\exp\{\theta_0+m(Y,X,\theta_1)\}.$$

Now write $f_{YX}(\cdot)$ as the joint density function of (Y, X) in the population. Then, with θ_0 and θ_1 denoting the true parameters,

$$\begin{aligned}\pi_d=\mathrm{pr}(D=d)&=\int H\{\theta_0+m(y,x,\theta_1)\}^d[1-H\{\theta_0+m(y,x,\theta_1)\}]^{1-d}f_{YX}(y,x)\mathrm{d}y\mathrm{d}x\\ &=\int[1-H\{\theta_0+m(y,x,\theta_1)\}]\exp[d\{\theta_0+m(y,x,\theta_1)\}]f_{YX}(y,x)\mathrm{d}y\mathrm{d}x.\end{aligned}$$

It then follows that the density of (Y, X) given D is

$$f_{YX|D=d}(y,x)=\frac{\exp[d\{\theta_0+m(y,x,\theta_1)\}]f_{YX}(y,x)}{[1+\exp\{\theta_0+m(y,x,\theta_1)\}]\pi_d}.$$

Recall that $\kappa = \theta_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$. Then equation (4) can now be computed as

$$\begin{aligned}&\mathrm{n}^{-1}\sum_{i=1}^{n}E[G\{R_i(\beta),X_i\}|D_i]\\ &=\sum_{d=0}^{1}\frac{n_d}{n\pi_d}\int G\{y-\mu(x,\beta),x\}\frac{\exp[d\{\theta_0+m(y,x,\theta_1)\}]}{1+\exp\{\theta_0+m(y,x,\theta_1)\}}f_{YX}(y,x)\mathrm{d}y\mathrm{d}x=\frac{n_0}{n\pi_0}\int\sum_{d=0}^{1}G\{y-\mu(x,\beta),x\}\frac{n_d/n_0}{\pi_d/\pi_0}\frac{\exp[d\{\theta_0+m(y,x,\theta_1)\}]}{1+\exp\{\theta_0+m(y,x,\theta_1)\}}f_{YX}(y,x)\mathrm{d}y\mathrm{d}x\\ &=\frac{n_0}{n\pi_0}\int G(r,x)\frac{1+\exp[\kappa+m\{r+\mu(x,\beta),x,\theta_1\}]}{1+\exp[\theta_0+m\{r+\mu(x,\beta),x,\theta_1\}]}f_{YX}\{r+\mu(x,\beta),x\}\mathrm{d}r\mathrm{d}x.\end{aligned}$$

The joint density of (Y, X) in the population is $f_{YX}(y, x) = f_{\varepsilon}\{y - \alpha_{\mathrm{true}} - \mu(x, \beta_{\mathrm{true}})\}\, f_X(x)$. Hence, $f_{YX}\{r + \mu(x, \beta), x\} = f_{\varepsilon}\{r - \alpha_{\mathrm{true}} - \tau(x, \beta, \beta_{\mathrm{true}})\}\, f_X(x)$. Thus,

$$
\begin{aligned}
&\mathrm{n}^{-1}\sum_{i=1}^{n} E[G\{R_i(\beta), X_i\}|D_i] \\
&= \frac{n_0}{n\pi_0}\int G(r, x)\frac{1+\exp[\kappa+m\{r+\mu(x,\beta_{\text{true}})-\tau(x,\beta,\beta_{\text{true}}),x,\theta_1\}]}{1+\exp[\theta_0+m\{r+\mu(x,\beta_{\text{true}})-\tau(x,\beta,\beta_{\text{true}}),x,\theta_1\}]} \\
&\times f_\varepsilon\{r-\alpha_{\text{true}}-\tau(x,\beta,\beta_{\text{true}})\}f_X(x)\mathrm{d}r\mathrm{d}x \\
&= \frac{n_0}{n\pi_0}\int G\{r \\
&\quad +\tau(x,\beta,\beta_{\text{true}}), x\}\frac{1+\exp[\kappa+m\{r+\mu(x,\beta_{\text{true}}),x,\theta_1\}]}{1+\exp[\theta_0+m\{r+\mu(x,\beta_{\text{true}}),x,\theta_1\}]} \quad \times f_\varepsilon(r \quad -\alpha_{\text{true}})f_X(x)\mathrm{d}r\mathrm{d}x.
\end{aligned}
$$

Now, since

$$\mathscr{K}(r, x, \beta_{\text{true}}, \Omega_{\text{true}}) = (1+\exp[\kappa+m\{r+\mu(x,\beta_{\text{true}}), x, \theta_1\}])(1+\exp[\theta_0+m\{r+\mu(x,\beta_{\text{true}}), x, \theta_1\}])^{-1},$$

we have that

$$n^{-1}\sum_{i=1}^{n} E[G\{R_i(\beta), X_i\}|D_i] = \frac{n_0}{n\pi_0}\int f_\varepsilon(r-\alpha_{\text{true}})f_X(x)\mathscr{K}(r, x, \beta_{\text{true}}, \Omega_{\text{true}})G\{r+\tau(x,\beta,\beta_{\text{true}}), x\}\mathrm{d}r\mathrm{d}x. \tag{17}$$

It follows from the convention in equation (4) and equation (17) that

$$
\begin{aligned}
&\frac{n\pi_0}{n_0}E\{Q_n(\alpha_{\text{true}},\beta,\Omega_{\text{true}})\} \\
&\quad = E\{Q_n(\alpha_{\text{true}},\beta,\Omega_{\text{true}})|D_1,\ldots,D_n\} \\
&\quad = n^{1/2}\int f_\varepsilon(r \\
&\quad -\alpha_{\text{true}})f_X(x)\mathscr{K}(r, x, \beta_{\text{true}}, \Omega_{\text{true}})\left[L\{r+\tau(x,\beta,\beta_{\text{true}}), x, \alpha(\beta,\Omega_{\text{true}}),\beta\} - \frac{\int L\{r+\tau(x,\beta,\beta_{\text{true}}), \upsilon, \alpha(\beta,\Omega_{\text{true}}),\beta\}\mathscr{K}\{r+\tau(x,\beta,\beta}{\int \mathscr{K}\{r+\tau(x,\beta,\beta_{\text{true}}), s,\beta,\Omega_{\text{true}}\}}\right.
\end{aligned}
$$

If $\beta = \beta_{\text{true}}$, since $\tau(x, \beta_{\text{true}}, \beta_{\text{true}}) = 0$, it follows directly that the last term is 0, and therefore $0 = E\{Q_n(\alpha_{\text{true}}, \beta_{\text{true}}, \Omega_{\text{true}})|D_1, \ldots, D_n\}$. Hence $Q_n(\alpha_{\text{true}}, \beta, \Omega_{\text{true}}) = 0$ is an unbiased estimating equation. If $\beta \neq \beta_{\text{true}}$, then in general we shall have $0 \neq \{Q_n(\alpha_{\text{true}}, \beta, \Omega_{\text{true}})|D_1, \ldots, D_n\}$.

Appendix B

Asymptotic theory

B.1. Notation and assumptions

In this section we introduce notation that is needed for our main theorem in Section 3.6, and we also state the formal assumptions under which this result will be valid.

Let $(\beta, \Omega) = \Theta$, and let Θ_{true} denote its true value. Recall equation (4), and define

$$c_* = \lim_{n\to\infty}(n_0/n),$$

$$\alpha(\beta,\Omega)=\frac{\sum_{d=0}^{1}(n_d/n)E\left(R(\beta)\left[\int f_X(x)\mathscr{K}\{R(\beta),x,\beta,\Omega\}\mathrm{d}x\right]^{-1}\middle|D=d\right)}{\sum_{d=0}^{1}(n_d/n)E\left(\left[\int f_X(x)\mathscr{K}\{R(\beta),x,\beta,\Omega\}\mathrm{d}x\right]^{-1}\middle|D=d\right)},$$

$$\mathscr{T}\{R(\beta),X,\Theta,f_X\}=L\{R(\beta),X,\alpha(\beta,\Omega),\beta\}-\frac{\int L\{R(\beta),x,\alpha(\beta,\Omega),\beta\}\mathscr{K}\{R(\beta),x,\Theta\}f_X(x)\mathrm{d}x}{\int\mathscr{K}\{R(\beta),x,\Theta\}f_X(x)\mathrm{d}x},$$

$$\mathscr{M}_\Omega=\sum_{d=0}^{1}c_*^{1-d}(1-c_*)^dE\left[\frac{\partial\mathscr{T}\{R(\beta_{\mathrm{true}}),X,\Theta,f_X\}}{\partial\Omega^{\mathrm{T}}}\,|D=d\right]\Bigg|_{\Theta=\Theta_{\mathrm{true}}},$$

$$\mathscr{M}_\beta=\sum_{d=0}^{1}c_*^{1-d}(1-c_*)^dE\left[\frac{\partial\mathscr{T}\{R(\beta),X,\Theta_{\mathrm{true}},f_X\}}{\partial\beta^{\mathrm{T}}}\,|D=d\right]\Bigg|_{\beta=\beta_{\mathrm{true}}}.$$

Define

$$G_{\mathrm{num}}(r,x,d,\Theta)=L\{r,x,\alpha(\beta,\Omega),\beta\}\tilde{\mathscr{K}}(r,x,d,\Theta),$$

$$G_{\mathrm{den}}(r,x,d,\Theta)=\tilde{\mathscr{K}}(r,x,d,\Theta),$$

$$\mathscr{A}_{\mathrm{num}}(r,\Theta)=\sum_{d=0}^{1}(n_d/n)E\{G_{\mathrm{num}}(r,X,D,\Theta)|D=d\},$$

$$\mathscr{A}_{\mathrm{den}}(r,\Theta)=\sum_{d=0}^{1}(n_d/n)E\{G_{\mathrm{den}}(r,X,D,\Theta)|D=d\}.$$

Write

$$\mathscr{H}_n(\beta,\Theta)=n^{-1/2}\sum_{i=1}^{n}\left[\frac{n^{-1}\sum_{j=1}^{n}G_{\mathrm{num}}\{R_i(\beta),X_j,D_j,\Theta\}}{n^{-1}\sum_{j=1}^{n}G_{\mathrm{den}}\{R_i(\beta),X_j,D_j,\Theta\}}-\frac{\mathscr{A}_{\mathrm{num}}\{R_i(\beta),\Theta\}}{\mathscr{A}_{\mathrm{den}}\{R_i(\beta),\Theta\}}\right]$$

and

$$W\{R_i(\beta), X_j, D_j, \Theta\}=\frac{G_{\text{num}}\{R_i(\beta), X_j, D_j, \Theta\} - \mathscr{A}_{\text{num}}\{R_i(\beta), \Theta\}}{\mathscr{A}_{\text{den}}\{R_i(\beta), \Theta\}}-\frac{\mathscr{A}_{\text{num}}\{R_i(\beta), \Theta\}[G_{\text{den}}\{R_i(\beta), X_j, D_j, \Theta\} - \mathscr{A}_{\text{den}}\{R_i(\beta), \Theta\}]}{\mathscr{A}^2_{\text{den}}\{R_i(\beta), \Theta\}}.$$

Also define

$$\tilde{Z}_i(\beta)=\{R_i(\beta), X_i, D_i\},$$

$$\tilde{z}=(r, x, d),$$

$$Q_1\{\tilde{Z}_i(\beta), \tilde{Z}_j(\beta), \Theta\}=W\{R_i(\beta), X_j, D_j, \Theta\}+W\{R_j(\beta), X_i, D_i, \Theta\},$$

$$Q_{2j}(\tilde{z}, \beta, \Theta)=E[W\{R(\beta), x, d, \Theta\}|D=j],$$

$$h_{1j}(\tilde{z}, \beta, \Theta)=E[Q_1\{\tilde{z}, \tilde{Z}(\beta), \Theta\}|D=j](j=0, 1),$$

$$h_2\{R_i(\beta), X_i, D_i, \Theta\}=\frac{n_0}{n}(1-D_i)h_{10}\{\tilde{Z}_i(\beta), \beta, \Theta\}+\frac{n_1}{n}D_i h_{11}\{\tilde{Z}_i(\beta), \beta, \Theta\}+\frac{n_0}{n}D_i Q_{20}\{\tilde{Z}_i(\beta), \beta, \Theta\}+\frac{n_1}{n}(1-D_i)Q_{21}\{\tilde{Z}_i(\beta), \beta, \Theta\},$$

$$m_{\theta_1}(y, x, \theta_1)=\frac{\partial m(y, x, \theta_1)}{\partial \theta_1},$$

$$\Phi(y, x, d, \Omega)=(1, m_{\theta_1}(y, x, \theta_1))^{\mathrm{T}}[d - H\{\kappa+m(y, x, \theta_1)\}],$$

$$\mathscr{N}_{\Omega}=-\sum_{d=0}^{1} c_*^{1-d}(1 - c_*)^d[E\{\partial\Phi(Y, X, D, \Omega)/\partial\Omega|D=d\}|_{\Omega=\Omega_{\text{true}}}]^{-1},$$

$$\begin{aligned}&\Lambda(Y_i, X_i, D_i, \Theta_{\text{true}})\\&\quad=\mathscr{M}_{\Omega}(\mathscr{N}_{\Omega}\Phi(Y_i, X_i, D_i, \Omega_{\text{true}}), \Psi(Y_i, X_i, D_i, \Omega_{\text{true}}))^{\mathrm{T}}\\&\quad - h_2\{R_i(\beta_{\text{true}}), X_i, D_i, \Theta_{\text{true}}\}+\mathscr{T}\{R_i(\beta_{\text{true}}), X_i, \Theta_{\text{true}}, f_X\},\end{aligned}$$

where the function $\Psi(Y_i, X_i, D_i, \Omega_{\text{true}})$ is defined in assumption 4 below. Finally, let

$$\Sigma=\sum_{d=0}^{1} c_*^{1-d}(1-c_*)^d \mathscr{M}_\beta^{-1}\mathrm{cov}\{\Lambda(Y,X,D,\Theta_{\mathrm{true}})|D=d\}(\mathscr{M}_\beta^{-1})^{\mathrm{T}}.$$

Next, introduce the following assumptions, under which the main result in Section 3.6 is valid.

Assumption 1. The error ε is independent of X. The error distribution F_ε is twice continuously differentiable, and the distribution F_X of X is once continuously differentiable. The corresponding densities are denoted by f_ε and f_X.

Assumption 2. There exists some $0 < c^* < 1$ such that $n_0/n \to c^*$.

Assumption 3. The function $\mu(x, \beta)$ is three times continuously differentiable with respect to β, $m(y, x, \theta_1)$ is twice continuously differentiable with respect to y and θ_1, and $\Phi(y, x, d, \Omega)$ is continuously differentiable with respect to Ω. Also, the matrices $\mathscr{M}_\beta$ and $E\{\partial\Phi(Y, X, D, \Omega)/\partial\Omega|D = d\}|_{\Omega=\Omega_{\mathrm{true}}}$ are invertible.

Assumption 4. The estimator $\hat{\theta}_0$ satisfies

$$\hat{\theta}_0-\theta_{0,\mathrm{true}}=n^{-1}\Psi(Y_i,X_i,D_i,\Omega_{\mathrm{true}})+o_p(n^{-1/2}),$$

for some function Ψ that satisfies $E\{\Psi(Y, X, D, \Omega_{\mathrm{true}})|D\} = 0$.

B.2. Proofs

We are now ready to give the proof of our main asymptotic result. Before giving a formal proof, let us first highlight the main steps of the proof. First, it follows from Appendix A.2 that $\hat{Q}_n(\alpha, \beta, \Omega)$ is an unbiased estimating function. Plugging in an estimator of α_{true}, we use a Taylor expansion of $\hat{Q}_{n,\mathrm{est}}(\hat{\beta}, \hat{\Omega}) = 0$ around the true β and Ω, which gives a regular asymptotically linear expansion of $n^{1/2}(\hat{\beta} - \beta_{\mathrm{true}})$. Finally we apply the central limit theorem to obtain the required asymptotic normality result. Along the way, we must show an asymptotic expansion for $\mathscr{H}_n(\beta, \Theta)$, which is given in lemma 1. The notation in the statement of this lemma was introduced in the previous section.

Lemma 1. Assume that assumptions 1–3 are valid. Then, for each β and Θ,

$$\mathscr{H}_n(\beta,\Theta)=n^{-1/2}\sum_{i=1}^{n} h_2\{R_i(\beta),X_i,D_i,\Theta\}+o_p(1),$$

where $E[h_2\{R(\beta), X, D, \Theta\}|D] = 0$.

Proof. Define

$$Z_{\mathrm{num}}\{R(\beta),\Theta\}=n^{-1/2}\sum_{j=1}^{n}[G_{\mathrm{num}}\{R(\beta),X_j,D_j,\Theta\}-\mathscr{A}_{\mathrm{num}}\{R(\beta),\Theta\}],$$

$$Z_{\mathrm{den}}\{R(\beta),\Theta\}=n^{-1/2}\sum_{j=1}^{n}[G_{\mathrm{den}}\{R(\beta),X_j,D_j,\Theta\}-\mathscr{A}_{\mathrm{den}}\{R(\beta),\Theta\}].$$

Since by assumption 2 we have that $n_1/n_0 \to c$, $0 < c < \infty$, it follows that $Z_{\mathrm{num}}\{R(\beta),\Theta\} = O_p(1)$ and $Z_{\mathrm{den}}\{R(\beta),\Theta\} = O_p(1)$, for each β and Θ. Hence, by a Taylor series expansion and assumption 3,

$$\frac{\mathrm{n}^{-1}\sum_{j=1}^{n}G_{\mathrm{num}}\{R(\beta),X_j,\Theta,D_j\}}{n^{-1}\sum_{j=1}^{n}G_{\mathrm{den}}\{R(\beta),X_j,\Theta,D_j\}}-\frac{\mathscr{A}_{\mathrm{num}}\{R(\beta),\Theta\}}{\mathscr{A}_{\mathrm{den}}\{R(\beta),\Theta\}}=\frac{\mathscr{A}_{\mathrm{num}}\{R(\beta),\Theta\}+n^{-1/2}Z_{\mathrm{num}}\{R(\beta),\Theta\}}{\mathscr{A}_{\mathrm{den}}\{R(\beta),\Theta\}+n^{-1/2}Z_{\mathrm{den}}\{R(\beta),\Theta\}}-\frac{\mathscr{A}_{\mathrm{num}}\{R(\beta),\Theta\}}{\mathscr{A}_{\mathrm{den}}\{R(\beta),\Theta\}}=\frac{n^{-1/2}Z_{\mathrm{num}}\{R(\beta),\Theta\}}{\mathscr{A}_{\mathrm{den}}\{R(\beta),\Theta\}}$$

Thus,

$$\mathrm{H}_n(\beta,\Theta)=n^{-3/2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{G_{\mathrm{num}}\{R_i(\beta),X_j,D_j,\Theta\}-\mathscr{A}_{\mathrm{num}}\{R_i(\beta),\Theta\}}{\mathscr{A}_{\mathrm{den}}\{R_i(\beta),\Theta\}}-\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\mathscr{A}_{\mathrm{num}}\{R_i(\beta),\Theta\}}{\mathscr{A}^2_{\mathrm{den}}\{R_i(\beta),\Theta\}}[G_{\mathrm{den}}\{R_i(\beta),X_j,D_j,\Theta\}-\mathscr{A}_{\mathrm{den}}\{R_i(\beta),\Theta\}]\right)+o_p(1)$$

By definition, $E\{\mathscr{B}_n(\beta,\Theta)|D_1,\ldots,D_n\}=0$. By the definition of $W\{R_i(\beta),X_j,D_j,\Theta\}$,

$$\mathscr{B}_n(\beta,\Theta)=n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}W\{R_i(\beta),X_j,D_j,\Theta\}.$$

Without loss of generality, we can make the first n_0 observations be the controls, and the last $n-n_0$ observations be the cases. Then,

$$\begin{aligned}
\mathrm{B}_n(\beta,\Theta)=&n^{-3/2}\sum_{i=1}^{n_0}\sum_{j=1}^{n_0}W\{R_i(\beta),X_j,D_j,\Theta\}\\
&+n^{-3/2}\sum_{i=n_0+1}^{n}\sum_{j=n_0+1}^{n}W\{R_i(\beta),X_j,D_j,\Theta\}\\
&+n^{-3/2}\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0}W\{R_i(\beta),X_j,D_j,\Theta\}\\
&+n^{-3/2}\sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n}W\{R_i(\beta),X_j,D_j,\Theta\}=n^{-3/2}\sum_{i=1}^{n_0}\sum_{j=1}^{i-1}Q_1\{\tilde{\mathrm{Z}}_i(\beta),\tilde{\mathrm{Z}}_j(\beta),\Theta\}\\
&+n^{-3/2}\sum_{i=n_0+1}^{n}\sum_{j=n_0+1}^{i-1}Q_1\{\tilde{\mathrm{Z}}_i(\beta),\tilde{\mathrm{Z}}_j(\beta),\Theta\}\\
&+n^{-3/2}\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0}W\{R_i(\beta),X_j,D_j,\Theta\}\\
&+n^{-3/2}\sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n}W\{R_i(\beta),X_j,D_j,\Theta\}+o_p(1).
\end{aligned}$$

An easy calculation shows that

$$\mathrm{var}\left[n^{-3/2}\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0}W\{R_i(\beta),X_j,D_j,\Theta\}-n_1n^{-3/2}\sum_{j=1}^{n_0}Q_{21}\{\tilde{\mathrm{Z}}_j(\beta),\beta,\Theta\}\right]\to 0,$$

and similarly

$$\mathrm{var}[n^{-3/2}\sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n} W\{R_i(\beta), X_j, D_j, \Theta\} - n_0 n^{-3/2}\sum_{j=n_0+1}^{n} Q_{20}\{\tilde{Z}_j(\beta), \beta, \Theta\} \to 0,$$

Hence we have shown that

$$\begin{aligned} \mathrm{B}_{n}(\beta,\Theta)=&\left(\tfrac{n_0}{n}\right)^{3/2} n_0^{-3/2}\sum_{i=1}^{n_0}\sum_{j=1}^{i-1} Q_1\{\tilde{Z}_i(\beta),\tilde{Z}_j(\beta),\Theta\}\\ &+\left(\tfrac{n_1}{n}\right)^{3/2} n_1^{-3/2}\sum_{i=n_0+1}^{n}\sum_{j=n_0+1}^{i-1} Q_1\{\tilde{Z}_i(\beta),\tilde{Z}_j(\beta),\Theta\}+n_1 n^{-3/2}\sum_{i=1}^{n_0} Q_{21}\{\tilde{Z}_i(\beta),\beta,\Theta\}\\ &+n_0 n^{-3/2}\sum_{i=n_0+1}^{n} Q_{20}\{\tilde{Z}_i(\beta),\beta,\Theta\}+o_p(1). \end{aligned}$$

Except for the factor $(n_0/n)^{3/2}$, the first term above is a classical symmetric U-statistic of order 2 applied to independent and identically distributed observations, since by convention the first n_0 observations are the controls. It then follows from standard U-statistic theory that (see, for example, Van der Vaart (1998))

$$\begin{aligned} \mathrm{B}_{n}(\beta,\Theta)=&\left(\tfrac{n_0}{n}\right)^{3/2} n_0^{-1/2}\sum_{i=1}^{n_0} h_{10}\{\tilde{Z}_i(\beta),\beta,\Theta\}\\ &+\left(\tfrac{n_1}{n}\right)^{3/2} n_1^{-1/2}\sum_{i=n_0+1}^{n} h_{11}\{\tilde{Z}_i(\beta),\beta,\Theta\}\\ &+n_1 n^{-3/2}\sum_{i=1}^{n_0} Q_{21}\{\tilde{Z}_i(\beta),\beta,\Theta\}+n_0 n^{-3/2}\sum_{i=n_0+1}^{n} Q_{20}\{\tilde{Z}_i(\beta),\beta,\Theta\}\\ &+o_p(1)\\ &=n^{-1/2}\sum_{i=1}^{n} h_2\{R_i(\beta),X_i,D_i,\Theta\}+o_p(1). \end{aligned}$$

This completes the proof.

B.2.1. Proof of theorem 1—Because of the unbiasedness of the estimating function (13) and the fact that expression (14) is consistent and asymptotically normally distributed for α_{true} when evaluated at $(\beta_{\mathrm{true}}, \Omega_{\mathrm{true}})$, the estimate is consistent for β_{true}, and $\alpha(\beta_{\mathrm{true}}, \Omega_{\mathrm{true}}) = \alpha_{\mathrm{true}}$. Set

$$\mathscr{J}(R(\beta),X,\beta,\Omega)=\mu_\beta(X,\beta)-\frac{\int \mu_\beta(x,\beta)\mathscr{K}\{R(\beta),x,\beta,\Omega\}f_X(x)\mathrm{d}x}{\int \mathscr{K}\{R(\beta),x,\beta,\Omega\}f_X(x)\mathrm{d}x},$$

$$c_{1n}(\beta,\Omega)=n^{-1}\sum_{i=1}^{n}\mathscr{J}\{R_i(\beta),X_i,\beta,\Omega\},$$

$$c_1(\beta,\Omega)=\sum_{d=0}^{1}\frac{n_d}{n}E[\mathscr{J}\{R(\beta),X,\beta,\Omega\}|D=d].$$

We use the fact that 0 = $\hat{Q}_{n,\mathrm{est}}(\beta, \hat{\Omega})|_{\beta=\hat{\beta}}$. By a Taylor series expansion and assumption 3,

$$0=\hat{Q}_{n,\mathrm{est}}(\beta_{\mathrm{true}},\Omega_{\mathrm{true}})+\frac{\partial}{\partial\beta^{\mathrm{T}}}\{n^{-1/2}\hat{Q}_{n,\mathrm{est}}(\beta_{\mathrm{true}},\Omega_{\mathrm{true}})\}n^{1/2}(\hat{\beta}-\beta_{\mathrm{true}})+\frac{\partial}{\partial\Omega^{\mathrm{T}}}\{n^{-1/2}\hat{Q}_{n,\mathrm{est}}(\beta_{\mathrm{true}},\Omega_{\mathrm{true}})\}n^{1/2}(\hat{\Omega}-\Omega_{\mathrm{true}})+o_p(1).$$

However, since $\hat{\alpha}(\beta_{\mathrm{true}}, \Omega_{\mathrm{true}})$ is a consistent estimator for α_{true}, it is clear that we have that

$$n^{-1/2}\{\partial\hat{Q}_{n,\mathrm{est}}(\beta,\Omega_{\mathrm{true}})/\partial\beta^{\mathrm{T}}\}_{\beta=\beta_{\mathrm{true}}}=\mathscr{M}_{\beta}+o_p(1)$$

and

$$n^{-1/2}\{\partial\hat{Q}_{n,\mathrm{est}}(\beta_{\mathrm{true}},\Omega)/\partial\Omega^{\mathrm{T}}\}_{\Omega=\Omega_{\mathrm{true}}}=\mathscr{M}_{\Omega}+o_p(1).$$

Hence it follows that

$$0=\hat{Q}_{n,\mathrm{est}}(\beta_{\mathrm{true}},\Omega_{\mathrm{true}})+\mathscr{M}_{\beta}n^{-1/2}(\hat{\beta}-\beta_{\mathrm{true}})+\mathscr{M}_{\Omega}n^{-1/2}(\hat{\Omega}-\Omega_{\mathrm{true}})+o_p(1).$$

Because of its form, another Taylor series expansion and under assumption 3,

$$\hat{Q}_{n,\mathrm{est}}(\beta_{\mathrm{true}},\Omega_{\mathrm{true}})=\hat{Q}_{n}(\alpha_{\mathrm{true}},\beta_{\mathrm{true}},\Omega_{\mathrm{true}})+c_1(\beta_{\mathrm{true}},\Omega_{\mathrm{true}})n^{1/2}\{\hat{\alpha}(\beta_{\mathrm{true}},\Omega_{\mathrm{true}})-\alpha(\beta_{\mathrm{true}},\Omega_{\mathrm{true}})\}+o_p(1).$$

However, we can obtain by the same argument as in Appendix A.2 that $c_1(\beta_{\mathrm{true}}, \Omega_{\mathrm{true}}) = 0$. In addition, using the same tools as in lemma 1, $n^{1/2}\{\hat{\alpha}(\beta_{\mathrm{true}}, \Omega_{\mathrm{true}}) - \alpha(\beta_{\mathrm{true}}, \Omega_{\mathrm{true}})\} = O_p(1)$. We have thus shown that

$$n^{1/2}(\hat{\beta}-\beta_{\mathrm{true}})=-\mathscr{M}_{\beta}^{-1}\{\hat{Q}_{n}(\alpha_{\mathrm{true}},\beta_{\mathrm{true}},\Omega_{\mathrm{true}})+\mathscr{M}_{\Omega}n^{1/2}(\hat{\Omega}-\Omega_{\mathrm{true}})\}+o_p(1). \quad (18)$$

Because (κ, θ_1) is estimated by ordinary logistic regression, and assumption 4 gives a representation for $\hat{\theta}_0 - \theta_{0,\mathrm{true}}$, it follows from standard theory that

$$n^{1/2}(\hat{\Omega}-\Omega_{\mathrm{true}})=n^{-1/2}\sum_{i=1}^{n}(\mathscr{N}_{\Omega}\Phi(Y_i,X_i,D_i,\Omega_{\mathrm{true}}),\Psi(Y_i,X_i,D_i,\Omega_{\mathrm{true}}))^{\mathrm{T}}+o_p(1).$$

We thus have from equation (18) that

$$n^{1/2}(\hat{\beta}-\beta_{\mathrm{true}})=\mathscr{M}_{\beta}^{-1}\left\{\hat{Q}_{n}(\alpha_{\mathrm{true}},\beta_{\mathrm{true}},\Omega_{\mathrm{true}})+\mathscr{M}_{\Omega}n^{-1/2}\sum_{i=1}^{n}(\mathscr{N}_{\Omega}\Phi(Y_i,X_i,D_i,\Omega_{\mathrm{true}}),\Psi(Y_i,X_i,D_i,\Omega_{\mathrm{true}}))^{\mathrm{T}}\right\}+o_p(1).$$

We can now apply lemma 1 to $\hat{Q}_n(\alpha_{\mathrm{true}}, \beta_{\mathrm{true}}, \Omega_{\mathrm{true}})$ with $G_{\mathrm{num}}(r, x, d, \Theta) = L\{r, x, \alpha(\beta, \Omega), \beta\}\,\tilde{\mathscr{K}}(r, x, d, \Theta)$ and $G_{\mathrm{den}}(r, x, d, \Theta) = \tilde{\mathscr{K}}(r, x, d, \Theta)$. Invoking lemma 1, it follows that

$$\hat{Q}_{n}(\alpha_{\mathrm{true}},\beta_{\mathrm{true}},\Omega_{\mathrm{true}})=n^{-1/2}\sum_{i=1}^{n}\mathscr{T}\{R_i(\beta_{\mathrm{true}}),X_i,\Theta_{\mathrm{true}},f_X\}-n^{-1/2}\sum_{i=1}^{n}h_2\{R_i(\beta_{\mathrm{true}}),X_i,D_i,\Theta_{\mathrm{true}}\}+o_p(1).$$

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

We have shown in Appendix A.2 that the first term has mean 0. Remember from lemma 1 that $E[h_2\{R(\beta_{\mathrm{true}}), X, D, \Theta_{\mathrm{true}}\}|D] = 0$. Moreover, the estimating equation for logistic regression is unbiased and assumption 4 ensures that $E[\Psi(Y, X, D, \Omega_{\mathrm{true}})|D] = 0$. Summarizing, we have shown that

$$\begin{aligned} &\mathrm{n}^{1/2}(\hat{\beta}-\beta_{\mathrm{true}})=\mathscr{M}_{\beta}^{-1}n^{-1/2}\sum_{i=1}^{n}\Lambda(Y_i,X_i,D_i,\Theta_{\mathrm{true}})\\ &+o_p(1),\Lambda(Y_i,X_i,D_i,\Theta_{\mathrm{true}})\\ &=\mathscr{M}_{\Omega}(\mathscr{N}_{\Omega}\Phi(Y_i,X_i,D_i,\Omega_{\mathrm{true}}),\Psi(Y_i,X_i,D_i,\Omega_{\mathrm{true}}))^{\mathrm{T}}\\ &-h_2\{R_i(\beta_{\mathrm{true}}),X_i,D_i,\Theta_{\mathrm{true}}\}+\mathscr{T}\{R_i(\beta_{\mathrm{true}}),X_i,\Theta_{\mathrm{true}},f_X\}, \end{aligned}$$

$$0=E\{\Lambda(Y,X,D,\Theta_{\mathrm{true}})|D\},$$

as claimed.

References

Ahn J, Albanes D, Berndt SI, Peters U, Chatterjee N, Freedman ND, Abnet CC, Huang WY, Kibel AS, Crawford ED, Weinstein SJ, Chanock SJ, Schatzkin A, Hayes RB. the Prostate, Lung, Colorectal and Ovarian Trial Project Team. Vitamin D-related genes, serum vitamin D concentrations and prostate cancer risk. Carcinogenesis. 2009; 30:769–776. [PubMed: 19255064]

Anderson, R. Modern Methods for Robust Regression. New York: Sage; 2008.

Babu GJ, Singh K. Inference on means using the bootstrap. Ann. Statist. 1983; 11:999–1003.

Buonaccorsi, JP. Measurement Eror: Models, Methods and Applications. Boca Raton: Chapman and Hall; 2010.

Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions. Biometrika. 2005; 92:399–418.

Chen Y-H, Carroll RJ, Chatterjee N. Retrospective analysis of haplotype-based case-control studies under a flexible model for gene-environment association. Biostatistics. 2008; 9:81–99. [PubMed: 17490987]

Chen Y-H, Chatterjee N, Carroll RJ. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. J. Am. Statist. Ass. 2009; 104:220–233.

Chen X, Linton O, Van Keilegom I. Estimation of semiparametric models when the criterion function is not smooth. Econometrica. 2003; 71:1591–1608.

Epstein M, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. Am. J. Hum. Genet. 2003; 73:1316–1329. [PubMed: 14631556]

Hall P, Horowitz J. Bootstrap critical values for tests based on generalized-method-of-moments estimators. Econometrica. 1996; 6:891–916.

Hu YJ, Lin DY, Zeng D. A general framework for studying genetic effects and gene–environment interactions with missing data. Biostatistics. 2010; 11:583–598. [PubMed: 20348396]

Huber, PJ. Robust Statistics. New York: Wiley; 1981.

Jiang Y, Scott AJ, Wild CJ. Secondary analysis of case-control data. Statist. Med. 2006; 25:1323–1339.

Kwee LC, Epstein MP, Manatunga AK, Duncan R, Allen AS, Satten GA. Simple methods for assessing haplotype-environment interactions in case-only and case-control studies. Genet. Epidem. 2007; 31:75–90.

Lele, S. Resampling using estimating equations. In: Godambe, UP., editor. Estimating Functions. New York: Oxford University Press; 1991. p. 295-304.

Li H, Gail MH, Berndt S, Chatterjee N. Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. Genet. Epidem. 2010; 34:427–433.

Lin DY, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies (with discussion). J. Am. Statist. Ass. 2006; 101:89–118.

Lin DY, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. Genet. Epidem. 2009; 33:256–265.

Modan MD, Hartge P, Hirsh-Yechezkel G, Chetrit A, Lubin F, Beller U, Ben-Baruch G, Fishman A, Menczer J, Struewing JP, Tucker MA, Wacholder S. for the National Israel Ovarian Cancer Study Group. Parity, oral contraceptives and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. New Engl. J. Med. 2001; 345:235–240. [PubMed: 11474660]

Monsees G, Tamimi R, Kraft P. Genomewide association scans for secondary traits using case-control samples. Genet. Epidem. 2009; 33:717–728.

Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. Statist. Med. 1994; 13:153–162.

Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika. 1979; 66:403–411.

Spinka C, Carroll RJ, Chatterjee N. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. Genet. Epidem. 2005; 29:108–127.

Van der Vaart, AW. Asymptotic Statistics. Cambridge: Cambridge University Press; 1998.

Wang CY, Wang S, Carroll RJ. Estimation in choice-based sampling with measurement error and bootstrap analysis. J. Econmetr. 1997; 77:65–86.

Zhao LP, Li SS, Khalid N. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am. J. Hum. Genet. 2003; 72:1231–1250. [PubMed: 12704570]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 1

Results of the simulation study with $n_1 = 500$ cases and $n_0 = 500$ controls, and a disease rate of approximately 1%[†]

	Results for normal model				Results for gamma model			
	Controls	**SPMLE**	**Robust**	**All**	**Controls**	**SPMLE**	**Robust**	**All**
$\theta_y = 0.00$								
Mean	0.992	0.991	1.001	0.992	1.002	1.005	1.003	1.003
sd	0.148	0.107	0.119	0.105	0.156	0.111	0.120	0.111
Est. sd	0.154	0.110	0.121	0.109	0.154	0.110	0.121	0.109
90%	0.917	0.911	0.918	0.912	0.892	0.897	0.899	0.901
95%	0.956	0.955	0.965	0.955	0.944	0.943	0.944	0.941
MSE Eff		1.898	1.537	1.965		1.963	1.665	1.957
$\theta_y = 0.25$								
Mean	0.999	1.001	0.990	1.078	1.001	0.997	0.993	1.120
sd	0.154	0.110	0.117	0.109	0.155	0.144	0.120	0.144
Est. sd	0.154	0.111	0.119	0.110	0.153	0.149	0.123	0.148
90%	0.911	0.905	0.908	0.818	0.900	0.924	0.901	0.797
95%	0.955	0.954	0.958	0.889	0.945	0.961	0.947	0.881
MSE Eff		1.951	1.720	1.303		1.148	1.643	0.680
$\theta_y = 0.50$								
Mean	0.995	0.994	0.989	1.177	0.986	0.848	1.024	1.297
sd	0.154	0.114	0.117	0.114	0.144	0.205	0.147	0.208
Est. sd	0.154	0.113	0.120	0.113	0.148	0.208	0.149	0.215
90%	0.903	0.898	0.904	0.525	0.906	0.818	0.905	0.587
95%	0.957	0.947	0.948	0.641	0.953	0.884	0.957	0.719
MSE Eff		1.822	1.704	0.531		0.323	0.938	0.159

[†] ‘Normal’ means that $\epsilon \sim N(0, 1)$, and ‘gamma’ means that ϵ is a centred and scaled gamma random variable with shape parameter 0.4. The analyses performed are ‘controls’ (using only controls), the semiparametric efficient method that assumes normality (‘SPMLE’), our new estimator (‘robust’), and ‘all’, which is the method that uses all the data while ignoring the case–control study. Over 1000 simulations, we computed the mean estimated β (‘mean’), its standard deviation (‘sd’), the mean estimated standard deviation (‘Est. sd’), the coverage for a nominal 90% confidence interval (‘90%’), the coverage for a nominal 95% confidence interval (‘95%’) and the mean-squared error efficiency (‘MSE Eff’) compared with using only the controls.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 2

Results of the simulation study described in Table 1, now with $n_1 = 150$ cases and $n_0 = 150$ controls[†]

	Results for normal model				Results for gamma model			
	Controls	**SPMLE**	**Robust**	**All**	**Controls**	**SPMLE**	**Robust**	**All**
$\theta_y = 0.00$								
Mean	0.991	0.993	1.005	0.992	0.998	0.991	1.019	0.990
sd	0.287	0.204	0.233	0.200	0.292	0.201	0.236	0.199
Est. sd	0.282	0.202	0.230	0.200	0.281	0.201	0.230	0.199
90%	0.891	0.908	0.910	0.905	0.892	0.900	0.916	0.902
95%	0.942	0.951	0.965	0.952	0.948	0.950	0.959	0.950
MSE Eff		1.973	1.509	2.043		2.103	1.526	2.151
$\theta_y = 0.25$								
Mean	1.008	1.016	0.983	1.092	1.007	0.994	0.974	1.118
sd	0.301	0.204	0.220	0.202	0.280	0.268	0.223	0.267
Est. sd	0.283	0.204	0.227	0.202	0.273	0.269	0.232	0.268
90%	0.874	0.893	0.933	0.867	0.903	0.900	0.928	0.864
95%	0.933	0.950	0.968	0.930	0.943	0.947	0.968	0.928
MSE Eff		2.156	1.856	1.834		1.088	1.551	0.921
$\theta_y = 0.50$								
Mean	0.986	0.987	0.974	1.173	0.985	0.837	1.006	1.292
sd	0.283	0.199	0.222	0.200	0.265	0.393	0.295	0.400
Est. sd	0.282	0.206	0.235	0.207	0.266	0.381	0.311	0.393
90%	0.903	0.918	0.936	0.798	0.900	0.864	0.938	0.808
95%	0.948	0.958	0.973	0.871	0.943	0.923	0.969	0.888
MSE Eff		2.003	1.597	1.143		0.388	0.806	0.287

[†] The disease rate is approximately 1%.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 3

Results of the vitamin D receptor data example in Section 6[†]

X	Results for our method			Results for controls only			Efficiency
	Estimate	**Lowe limitr**	**Upper limit**	**Estimate**	**Lowe limitr**	**Upper limit**	
SNP 1	0.015	−0.165	0.195	−0.029	−0.262	0.204	1.68
SNP 2	0.023	−0.047	0.093	0.039	−0.069	0.146	2.36
SNP 3	0.015	−0.062	0.092	−0.045	−0.161	0.070	2.25

[†] Three analyses are displayed, one each when X is SNP 1, SNP 2 and SNP 3. Displayed are the parameter estimates of the slope for X (‘estimate’), and lower (‘lower’) and upper (‘upper’) 95% confidence intervals. Our method is contrasted with using linear regression among the controls only. Also displayed is the ‘efficiency’, which is defined as the square of the ratio of the lengths of the confidence intervals.